A Birds Eye View on System Identification

Manfred Deistler

Abstract

System identification is concerned with obtaining good models from data, i.e. with data driven modeling. In this contribution the aim is to explain and discuss ideas, general approaches and theories underlying identification of linear systems. Identification of linear systems is a nonlinear problem and is prototypical also for many parts of identification of nonlinear systems.

1 Introduction

The art of identification is to find a good model from, in general, noisy data. This is an important problem in many areas of application. Often the task of identification is so complex, that it cannot be performed with the naked eye and systematic approaches have to be used. This is done, partly under quite different perspectives, in statistics, econometrics, system theory and the field of inverse problems.

The main steps in identifications are:

- Specification of the model class, i.e. of the class of all a priori feasible candidate systems. In this step the a priori information concerning the phenomenon to be modeled is incorporated. This typically includes, for instance, the selection of (candidates for) the input-variables or assumptions on the relation between the variables.
- Specification of the class of observations, data preprocessing.
- Identification in the narrow sense: An identification procedure is a rule, in the automatic case a function, attaching a system from the model class to the data. In this step the emphasis is on the development of procedures and algorithms on one side and on their evaluation on the other side.

Here only identification from equally spaced, discrete time, time series data $y_t = (y_t^{(i)})^{i=1\dots s} \in \mathbb{R}$, $t = 1 \dots T$ is considered. For explanation of time series data, dynamic systems are often natural candidates.

In this contribution the focus is on what we call the main stream theory for identification of linear systems (see [6], [7]). We add a few remarks on alternative model classes and approaches for identification of linear systems and on identification of nonlinear systems.

The mainstream theory of identification deals with the following setting:

- The model class consists of linear, time-invariant, finite dimensional, causal and stable systems only. The classification of the variables into inputs and outputs is given a priori.
- Uncertainty is modeled by the use of stochastic models for noise. In particular here the noise is assumed to be stationary with a rational spectral density. These assumptions on the noise are in a sense standard, but are nevertheless not innocent. The have been critisized on grounds of not being justified in a number of applications (see e.g. [25]). In our opinion, stochastic noise models are at least an important "test bed" for evaluating identification procedures.
- The observed inputs are free of noise and uncorrelated with the noise process.
- The approach to estimation is semi-nonparametric in the following sense: In general the parameterspace for describing system- and noise parameters will be not finite-dimensional, since e.g. systems of arbitrarely high orders are considered. In this approach the model class is broken down into subclasses such that each subclass has a finite- dimensional parameter-space. Estimation then consists of two steps: The model selection step, where the subclass is estimated by a vector of integers, characterizing this subclass. Once the subclass is obtained, its parameter-space is a subset of a suitable Euclidian space and estimation is concerned with estimating a parameter, which is a vector of real-valued entries, in this space.
- For the statistical analysis, emphasis is laid on asymptotic properties (consistency, asymptotic normality and asymptotic efficiency), mainly because finite sample properties are hard to obtain analytically.

We consider the following three "modules" in the theory of system identification:

- Structure theory: Here an idealized problem is considered, as we commence from the stochastic processes generating the data or their population second moments rather than from the data. In the ergodic case one could also say that we commence from an infinite, rather than from a finite data string. The relation between "external behavior" (as described e.g. by the population second moments of the observations) and "internal" (system and noise-) parameters is analysed. Identifiabilty, realization- and parametrization theory are important parts of structure theory.
- Estimation of real-valued parameters for a given subclass: Here we commence from a given subclass whose parameter space is a subset of an Euclidean space and in addition contains a non-void open subset of this space. Estimators are often found from general principles, here in particular from optimizing a likelihood-type criterion function over the parameter-space.

- Model selection: In general the orders or the relevant inputs are not known a priori and have to be determined from the data. One way of doing this is e.g. estimation of integers characterizing the orders by information criteria like AIC or BIC, or, more generally by using a criterion defining a trade-off between the quality of fit to the data achievable in a certain model-subclass and the complexity of this subclass

2 Structure Theory

As has been stated already, structure theory is concerned with an analysis of the relation between external behavior and internal parameters. Such an analysis turns out to be important for a deeper understanding of many identification procedures. For the linear mainstream case, the relation between the population second moments of the observations or equivalently the transfer-functions (and noise covariance matrices) and the system (and noise) parameters is considered.

Main model classes for linear systems are:

- $AR(X)$ models
- $ARMA(X)$ models
- State space models

In many applications $AR(X)$ systems still dominate for a number of reasons. Main advantages of (unrestricted), $AR(X)$ models are:

- There are no problems of non-identifiability; in more general terms structure theory is so simple, that for standard situations it does not have to be considered separately.
- Least squares estimators are of maximum likelihood type; they are explicitly given, fast to calculate and asymptotically efficient.

Things are different in case of "structural" a priori restrictions (i.e. if restrictions on the parameter space are imposed by a priori knowledge); but nevertheless, also then $AR(X)$ system identification is "easier" compared to the $ARMA(X)$ or state space case.

On the other hand $AR(X)$ models are less flexible than $ARMA(X)$ and state space models in the sense that, in general, more parameters are needed to achieve the same quality of approximation.

In this contribution, we will mainly consider the case where we have no observed inputs.

Here the focus is on state space systems, but we also consider the $ARMA(X)$ case. A state space system in innovations representation is of the form

$$x_{t+1} = Ax_t + B\epsilon_t(+Lz_t) \tag{1}$$

$$y_t = Cx_t + \epsilon_t(+Dz_t) \tag{2}$$

where y_t are the s-dimensional outputs, x_t is the n-dimensional state, (ϵ_t) is, in general unobserved, s-dimensional white noise (i.e. $E\epsilon_t = 0, E\epsilon_s\epsilon_t^{'} = \delta_{st}\Sigma$, where E denotes expectation and δ_{st} is the Kronecker symbol) and z_t are the m-dimensional observed inputs. The random variables y_t, x_t, ϵ_t and z_t are defined over an underlying probability space $(\Omega, \mathcal{A}, \mathcal{P})$. $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times s}$, $L \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{s\times n}$ and $D \in \mathbb{R}^{s\times m}$ are parameter matrices.

Throughout we assume that the *stability condition*

$$|\lambda_{max}(A)| < 1 \tag{3}$$

where λ_{max} denotes an eigenvalue of maximum modulus, and the *miniphase condition*

$$|\lambda_{max}(A - BC)| \leq 1 \tag{4}$$

hold. The steady state solution of (1) (2) is given by

$$y_t = C(Iz^{-1} - A)^{-1}(B\epsilon_t(+Lz_t)) + \epsilon_t(+Dz_t) \tag{5}$$

Here z is used for a complex variable as well as for the backward shift on the integers $\mathbb{Z}$, i.e. $z(y_t|t \in \mathbb{Z}) = (y_{t-1}|t \in \mathbb{Z})$.

In addition, throughout we assume that $Ez_s\epsilon_t^{'} = 0$ holds and that Σ is non-singular.

$ARMA(X)$ systems are (vector-) difference equations of the form

$$a(z)y_t = b(z)\epsilon_t(+d(z)z_t) \tag{6}$$

where

$$a(z) = \Sigma_{j=0}^{p} a_j z^j \quad ; \quad b(z) = \Sigma_{j=0}^{q} b_j z^j \quad ; \quad d(z) = \Sigma_{j=0}^{r} d_j z^j \quad ;$$

$$a_j; b_j \in \mathbb{R}^{s\times s} \quad ; \quad d_j \in \mathbb{R}^{s\times m}$$

We assume that the *stability condition*

$$\det a(z) \neq 0 \qquad |z| \leq 1 \tag{7}$$

and the *miniphase condition*

$$\det b(z) \neq 0 \qquad |z| < 1 \tag{8}$$

hold, and again we assume

$$Ez_s\epsilon_t^{'} = 0$$

and that Σ is nonsingular. The steady state solution then is given by

$$y_t = a^{-1}(z)[b(z)\epsilon_t(+d(z)z_t)] \tag{9}$$

Note that by (3) or (7) (and by assuming stationarity for (z_t)) the infinite sums in (5) and (9) respectively, i.e.

$$y_t = \Sigma_{j=0}^{\infty} k_j \epsilon_{t-j} (+\Sigma_{j=0}^{\infty} l_j z_{t-j}) \tag{10}$$

where, e.g,

$$k_j = CA^{j-1}B \quad , \quad j > 0 \quad , \quad k_0 = I \tag{11}$$

and

$$k(z) = \Sigma_{j=0}^{\infty} k_j z^j = a^{-1}(z)b(z) \tag{12}$$

converge e.g. in the mean squares sense. In addition (y_t) and (x_t) are stationary processes.

From now onwards, we will, for the sake of brevity of notation, unless the contrary is stated explicitly, restrict ourselves to the case, where there are no observed inputs. Then the external behavior of (1), (2) or (6) is described by the covariance function $\gamma : \mathbb{Z} \to \mathbb{R}^{s\times s}$, $\gamma(t) = Ey_t y_0'$ of the process (y_t) or equivalently by its spectral density $f : [-\pi, \pi] \to \mathbb{C}^{s\times s}$ defined by

$$f(\lambda) = (2\pi)^{-1}\Sigma_{t=-\infty}^{\infty} e^{-i\lambda t}\gamma(t) \tag{13}$$

From (10), f is given by

$$f(\lambda) = (2\pi)^{-1}k(e^{-i\lambda})\Sigma k^*(e^{-i\lambda}) \tag{14}$$

where $*$ denotes the conjugate transpose. Throughout we assume $k(0) = I$. This implies no restriction for f and establishes a one-to-one relation between f and (k, Σ).

Under our assumptions,

- Every state space system (1) (2) and every $ARMA$ system (6) has a rational transfer function $k(z)$ which is analytic in a disk containing the closed unit disk (and thus is causal and stable) and which satisfies $\det k(z) \neq 0$, $|z| < 1$.
- Conversely, for every rational transfer function $k(z)$ which is analytic in a disk containing the closed unit disk and which satisfies $\det k(z) \neq 0$, $|z| < 1$ and $k(0) = I$ there is a stable and miniphase state space -, and a stable and miniphase $ARMA$ representation.

Thus, in particular, state space- and $ARMA$ representations are two alternative ways to describe the same class of external (input/output) behaviors $k(z)$. Note that the assumption $k(0) = I$ is a normalizing condition defining Σ. We have ([14], chapter 1):

Any rational and a.e. nonsingular spectral density matrix f may be uniquely factorized as in (14) where $k(z)$ is rational, analytic whithin a circle containing the closed unit disk, $\det k(z) \neq 0$, $|z| < 1$ and $k(0) = I$ and where $\Sigma > 0$.

Consider the following set of $s \times s$ transfer functions: $U_A = \{k|k$ is rational, $k(0) = I$, $k(z)$ has no poles for $|z| \leq 1$ and no zeros for $|z| < 1\}$. By $M(n) \subset U_A$ we denote the set of all transfer functions of order n (to be more precise, the set of all transfer functions corresponding to minimal state space systems with state dimension n). By T_A we denote the set of all triples (A, B, C), where s is fixed but n is arbitrary, satisfying (3) and (4), by $S(n) \subset T_A$ the subset of all (A, B, C) for fixed n and by $S_m(n) \subset S(n)$ the subset of all minimal (A, B, C). We define the mapping $\pi : T_A \to U_A$ such that $\pi(A, B, C) = C(Iz^{-1} - A)^{-1}B + I$ (also defined by (11)).

Now, T_A is not a "good" parameter space because:

- T_A is not finite dimensional
- π is (surjective but) not injective, i.e. we do not have identifiability
- There exists no continuous selection, in the sense that there is no continuous mapping attaching to every $k \in U_A$ a unique element from the equivalence class $\pi^{-1}(k)$

Here U_A is endowed with the socalled pointwise topology T_{pt} [14] which corresponds to the relative topology in the product space $(\mathbb{R}^{s \times s})^{\mathbb{N}}$ for the coefficients $(k_j | j \in \mathbb{N})$.

In order to obtain "good" parameter spaces, U_A and T_A are broken into bits, U_α and T_α say, $\alpha \in I$ such that

- π restricted to T_α, $\pi/T_\alpha : T_\alpha \to U_\alpha$ is bijective. Injectivity of π/T_α implies identifiability
- U_α is finite dimensional in the sense that $U_\alpha \subset \cup_{i=1}^{n} M(i)$ for some n. Usually, taking into account the restrictions in T_α, T_α is reparametrized by expressing the $(A, B, C) \in T_\alpha$ by their "free" parameters, τ say. We use T_α also for this set of free parameters τ and we assume that this T_α contains an open set in an embedding Euclidian space $\mathbb{R}^{d_\alpha}$. The mapping $\Psi_\alpha : U_\alpha \to T_\alpha$: $\Psi_\alpha(\pi(\tau)) = \tau \ \forall \tau \in T_\alpha$ is called a *parametrization*.
- The parametrization $\Psi_\alpha : U_\alpha \to T_\alpha$ is a homeomorphism; this is an assumption of well-posedness

- U_α is T_{pt}-open in its closure $\bar{U}_\alpha$
- $\cup_{\alpha \in I}\, U_\alpha$ is a cover of U_A

Usually, I is a set of vectors of integers (multiindices) characterizing the bits U_α and T_α. Note that not all approaches used have the desirable properties listed above.

Completely analogous statements hold for the $ARMA$ case, where (using the same symbols) the mapping π is defined by $\pi(a, b) = a^{-1}b$.

The most common approaches are:

- *Canonical forms* defining decompositions of $M(n)$, such as *echelon forms* [14] or *balanced realizations*. Here $M(n)$ is decomposed into sets U_α of different dimension. Echelon forms for state space and $ARMA$ systems have "nice" free parameters in terms of elements of (A, B, C), and of (a, b) and define a very simple bijection between state space and $ARMA$ parameters. Balanced realizations (which only exist for state space systems) have "nice" parameter spaces, but the free parameters are rather complicated transformations of the elements of (A, B, C).
- The overlapping description of the manifold $M(n)$ by local coordinates ([14]).
- The "full parametrization" for state space systems. Here $S(n) \subset \mathbb{R}^{n^2+2ns}$ or $S_m(n)$ are used as parameter spaces for $\bar{M}(n)$ (the closure of $M(n)$ in U_A) or $M(n)$ respectively. Clearly in this case we do not have identifiability. For $k \in M(n)$, the classes of observationally equivalent (A, B, C), $\pi^{-1}(k) \cap S(n)$ are n^2-dimensional manifolds.
- Data driven local coordinates, $DDLC$, for state space systems. Here $S_m(n)$ is reparamized in terms of coordinates that separately describe the tangent space to the manifold of observationally equivalent (minimal) systems corresponding to an initial estimator at a suitably chosen point and its $2ns$-dimensional orthocomplement [18], [20]. The orthocomplement then is taken as the new parameter space.
- $ARMA$ systems with prescribed column-degrees ([5])
- $ARMA$ parametrizations commencing from writing k as $c^{-1}p$ where c is a least common denominator polynomial for k and where the degrees of c and p serve as integer valued parameters.

3 Estimation for a Given Subclass

Here we commence from the data $y_t, t = 1 \ldots T$ and we assume that U_α is given. We in addition assume that we have identifiability and that the parametrization

$\Psi_\alpha : U_\alpha \to T_\alpha$ has the desirable properties listed above.

Let $\tau \in T_\alpha \subset \mathbb{R}^{d_\alpha}$ denote the vector of free parameters for U_α and let $\sigma \in \underline{\Sigma} \subset \mathbb{R}^{\frac{n(n+1)}{2}}$ denote the vector formed by the on and above diagnonal elements of Σ. $\underline{\Sigma}$ corresponds to the set of symmetric positive definite matrices. We assume that the overall parameter space is of the form $\Theta = T_\alpha \times \underline{\Sigma}$.

Many identification procedures, at least asymptotically, commence from the sample second moments of the observations:

$$\hat{\gamma}(s) = T^{-1}\Sigma_{t=1}^{T-s} y_{t+s} y_t^{'}, \qquad s \geq 0$$

Now, $\hat{\gamma}$ can be directly realized as an MA system, "typically" of order $T.s$. By $\check{k}_T$ we denote the corresponding transfer function. Clearly, in many cases, its order is too high. "Typical" identification procedures therefore consist of two steps:

- A "projection" or model reduction step, where $\check{k}_T$ is approximated by an element $\hat{k}_T$ say, in U_α ($\bar{U}_\alpha$). From a statistical point of view, this is the essential information concentration step and the statistical properties depend on the way the approximation is defined.
- A realization step, where $\hat{k}_T \in U_\alpha$ is realized by $\tau \in T_\alpha$. This step is important for a number of reasons, for instance from a numerical point of view, however certain statistical properties do not depend on this step.

One may distinguish between two types of estimation procedures, namely:

- Optimization based procedures (M-estimators), which are obtained from optimizing a criterion function over the parameterspace and where the estimators are not given explicitly

and

- Direct procedures, such as instrumental variable methods or subspace methods, where the estimators are explicit functions of the data

The most common criterion function is the Gaussian (log) likelihood function, which (when multiplied by $-2T^{-1}$) is (up to a constant) of the form

$$\hat{L}_T(\theta) = T^{-1} logdet\Gamma_T(\theta) + T^{-1} y'(T)\Gamma_T(\theta)^{-1} y(T) \tag{15}$$

where $y(T) = (y_1', \ldots, y_T')'$ is the stacked vector of observations, $\theta = (\tau', \sigma')'$ is the vector of system and noise parameters, $y(T;\theta)$ is a stacked vector of random variables formed from the outputs of systems with system and noise parameters θ, (in an analogous way as $y(T)$) and finally $\Gamma_T(\theta) = Ey(T;\theta)y(T;\theta)'$. The Gaussian maximum likelihood estimator (MLE) then is defined by

$$\hat{\theta}_T = argmin_{\theta\in\Theta} \hat{L}_T(\theta)$$

It is well known that, although the likelihood is written down as if the observations were Gaussian the asymptotic properties of the MLE do not depend on the Gaussianity of the observations. In addition, for the likelihood (15), the Gaussian distribution is assumed to come from a stationary process; transients in the observations do not influence the asymptotic properties of the MLE.

There exist a number of alternative criterion functions such as the Whittle Likelihood of Ljung's prediction error criterion [17] which (in most cases) give asymptotically equivalent estimators. The Whittle likelihood is of the form:

$$\hat{L}_{w,T}(k,\sigma) = logdet\Sigma + (2\pi)^{-1}\int_{-\pi}^{\pi} tr[(k(e^{-i\lambda})\Sigma k^*(e^{-i\lambda}))^{-1} I(\lambda)]d\lambda \qquad (16)$$

where tr denotes the trace and I is the periodogramm, i.e. the Fourier transform of $\hat{\gamma}$. Formula (16) shows the approximation of I by $k \in U_\alpha$ in a clear way.

In maximum likelihood estimation a number of observations are important:

- For "natural" parameter spaces, the likelihood function is not necessarely semi-continuous and thus the existence of its optimum is not guaranteed (see [10]).
- In general, the MLE is not given by an explicit function of the data; thus the estimators are obtained by a numerical optimization procedure.
- $\hat{L}_T$ depends on τ only via the corresponding transfer function k, thus (with a slight sloppyness in notation) we may define a "coordinate-free" likelihood function $\hat{L}_T(k,\sigma)$.
- Neither T_α nor U_α are closed sets and boundary points may occur in optimizing the likelihood function (see [14]).

As far as consistency of the MLE's is concerned, the first correct proofs have been given in [12] (for the univariate case) and [11], [8], for a general result see also [14]. Coordinate free consistency says that for $k_0 \in \bar{U}_\alpha$ (where k_0 denotes the true system) and if $limT^{-1}\Sigma_{t=1}^{T-s}\varepsilon_{t+s}\varepsilon_t^{'} = \delta_{0,s}\Sigma$ a.e. we have for the MLE's $\hat{k}_T \to k_0$ a.e. and $\hat{\Sigma}_T \to \Sigma_0$ a.e. The proof uses the basic idea of [24] developed for the i.i.d. case. The specific additional difficulties are not only due to the fact that the observations are dependent, but also due to the fact that the "natural" parameter spaces are not compact. As can be shown

$$\begin{aligned} lim_{T\to\infty}\hat{L}_T(k,\sigma) = L(k,\sigma) = \\ logdet\Sigma + (2\pi)^{-1}\int_{-\pi}^{\pi} tr[(k(e^{-i\lambda})\Sigma k^*(e^{-i\lambda}))^{-1} \\ (k_0(e^{-i\lambda})\Sigma_0 k_0^*(e^{-i\lambda}))]d\lambda \quad a.e. \end{aligned} \qquad (17)$$

holds, where the subscript 0 again denotes the true quantities and the asymptotic likelihood L has a unique minimum at k_0, Σ_0. (Note the similarity of (16) and L). Clearly pointwise convergence in (17) alone does not ensure convergence of the optima, but in our case, the latter can be shown, in particular since $(\hat{k}_T, \hat{\Sigma}_T)$ can be shown to enter a compact set. For a detailed proof of the consistency result, see [14], chapter 4.

Even for $k_0 \notin \bar{U}_\alpha$, the MLE's have a "generalized consistency" property, as they converge a.e. to the set D of minimizes of L over $\bar{U}_\alpha \times \underline{\Sigma}$.

Now consistency "in coordinates", i.e. for the parameter estimators $\hat{\tau}_T = \psi_\alpha(\hat{k}_T)$ follows directly from the consistency of $\hat{k}_T$ and the continuity of ψ_α (and from the openness of U_α in $\bar{U}_\alpha$), if $\tau_0 \in T_\alpha$ holds.

Under additional assumptions (see [14]) asymptotic normality can be shown by using the idea explained in [4], extended to the stationary case.

For actual calculation of estimators, the usual procedure consists of a consistent explicit estimator, e.g. a subspace procedure, to obtain an initial estimator in the first step and one Gauß-Newton step is order to obtain an asymptotically efficient estimator.

If T_β denotes a parameterspace obtained by a diffeomorphic mapping from T_α, then the transformation of the asymptotic distributions of the MLE's is straightforward; nevertheless the choice of parameterspaces turns of to be important from a numerical point of view, where it is taken into account that optimization has to be performed over a grid.

Explicit estimation procedures are usually numerically fast and reliable, but are in many cases not asymptotically efficient. Recently so called subspace identification procedures [1], [15], [23] have attracted a lot of attention. Subspace estimators are for state space systems and they are based on a realization algorithm combined with a model reduction step. Usually the model reduction step is performed by omitting the smaller eigenvalues in a singular value decomposition. For the case of observed inputs, subspace procedures turn out be more intricate. The statistical properties of classes subspace procedures have been investigated e.g. in [9], [3] and [2]

4 Model Selection

Here we confine our discussion to the problem of estimating the model order n. In many cases information criteria defining a trade-off between the quality of fit achievable in a certain model-subclass and the complexity of this subclasses are used for this purpose. Note that we here have a situation which is "closure nested", i.e. $n_1 < n_2$ implies $\bar{M}(n_1) \subset \bar{M}(n_2)$ and the dimension of $M(n_1)$ is

smaller than the dimension of $M(n_2)$. In particular criteria of the form

$$A(n) = logdet\hat{\Sigma}_T(n) + 2ns.c(T)T^{-1} \tag{18}$$

where $\hat{\Sigma}_T(n)$ is the MLE of Σ_0 over $\bar{M}(n) \times \underline{\Sigma}$ and $c(T)$ is a prescribed positive function of T, are frequently used. For $c(T) = 2$ we obtain the AIC, for $c(T) = c.logT$, $c \geq 1$, the BIC criterium. The corresponding order estimate $\hat{n}_T$ is obtained by minimizing $A(n)$ (in a certain range of integers).

The statistical properties of such estimators have been analysed by Hannan [13], see also [14], chapter 5. In particular (under suitable additional assumptions) $\hat{n}_T$ is (strongly) consistent if

$$lim_{T\to\infty}\frac{c(T)}{T} = 0$$

and

$$liminf_{T\to\infty}\frac{c(T)}{loglogT} > 0$$

hold (and thus BIC gives consistent estimators) and AIC does not give consistent estimators, and asymptotically leads to overestimation of the true order n.

Taking a closer look, things turn out to be more subtle. One may argue, as has been done, e.g. in [22], that in most cases order estimation is only an intermediate goal. Shibata showed that under certain assumptions, in particular if the true system is of infinite order, an autoregressive spectral estimate based on AIC and MLE is asymptotically optimal.

Here we concentrate on two issues. The first one may be entitled "decomposition into subclass is in the eye of the beholder". Consider e.g. AR models for $s = 1$ of the form

$$y_t + a_1 y_{t-1} + a_2 y_{t-2} = \epsilon_t$$

Then, considering only the system parameters, "usually" we have the following parameterspaces:

$T = \{(a_1, a_2) \in \mathbb{R}^2 | 1 + a_1 z + a_2 z^2 \neq 0 | z| \leq 1\}$
$T_0 = \{(0, 0)\}$
$T_1 = \{(a_1, 0) || a_1| < 1, a_1, a_1 \neq 0\}$
$T_2 = T - (T_0 \cup T_1)$
Now, from the Bayesian deriviation of BIC [21] we see that, in order to obtain BIC, T_0, T_1 and T_2 must have strictly positive prior probabilities; thus BIC has to be justified by some kind of a priori knowledge. For instance a flat prior on T would suggest just to use the MLE over T, rather than to do model selection in a first step. In addition other prior distributions may give positive prior probabilities to other low dimensional subsets of T and thus result in an other decomposition of T.

The second issue is concerned with properties of post model selection estimators, i.e. with properties of estimators for real-valued parameters taking into account the uncertainty coming from model selection. One may argue, that, if a (strongly) consistent model selection criterion is used, then the true model order is known from a certain sample size onwards and thus the asymptotic variance of the estimators for τ after model selection is the same as in the case where the true order is a priori known. As has been shown in [16] this argument is grossly misleading, because it is pointwise in the parameterspace and does not hold uniformly there.

5 Linear Non-Mainstream Cases

In a number of important cases, the systems are linear, but the models or their identification is not in the mainstream setting. We do not intend to give a survey on such cases here, but we only make a few remarks. Important special cases are:

- Linear systems with time-varying parameters. There several different approaches to this problem, such as systems with slowly varying parameters, where the variation of coefficients is described by an autoregression, or systems with structural changes, which may be triggered by a random mechanism, such as Markov switching models, or smooth transition models.
- Identification in closed loop.
- The wide area of symmetric modelling, where no a priori distinction between inputs and outputs is made, errors-in-variables and linear dynamic factor models; the latter are used in particular for high dimensional time series.
- Unstable systems, in particular integration and cointegration, which is of great importance for econometrics.
- Long memory

6 Nonlinear Systems

Of course there is a wide range of nonlinear systems and identification of nonlinear systems is a word like "non-elephant zoology" (also in the sense that linear systems are "huge animals"). Again, as in section 5, we do not intend to give a survey on this topic, we only make a few remarks on this field. Identification of nonlinear systems consists of a number of only weakly connected subareas. The most important of these subareas are:

- The asymptotic theory for M-estimation for parametric classes of nonlinear dynamic systems, see e.g. [19]. Identification of linear systems is a nonlinear problem, since the mapping from data to parameters is nonlinear. Identification of nonlinear systems uses partly the same ideas as identification of linear systems. The main problem in the setting of nonlinear systems is, that there is no general structure theory available, and thus, for instance identifiability is often assumed rather than shown.

- Neural nets are a frequently used model class also for time series. Recurrent neural nets are a particular class of dynamic nonlinear models. Identification of neural nets is a semi-parametric problem.

- Nonparametric estimation for nonlinear time series models, e.g. estimation of nonlinear autoregressions

$$y_t = g(y_{t-1}, \ldots, y_{t-p}) + \epsilon_t \tag{19}$$

  by kernel methods is, an area which has received substantial attention in the last two decades. The systems (19) can be generalized by replacing ϵ_t by a model for the conditional variance of y_t. Another important class in this area are socalled additive models, which are $MISO$ models, where the effects of the single inputs are nonlinear but additive, i.e. there is no interaction effect of different single inputs.

  Of central interest for nonparametric estimation is the choice of design parameters, such as the bandwith of a kernel, and, in asymptotic theory e.g. the rates of convergence.

- Special classes of nonlinear systems, justified by "physical" a priori knowledge or "stylized facts" in data, such as $GARCH$-type models or stochastic volatility models for explaining or forecasting conditional variances, in particular for finance data, have attracted a substantial number of researchers recently.

- Chaotic systems and their identification have been considered in the last 25 years, but the number of convincing success stories in applications seems to be limited.

7 Present State and Future Developments

Theory and methods in system identification have reached a certain state of maturity. There is a large body of methods and theory available serving the needs for many applications, but nevertheless, in many cases, identification is still not a standard task and needs a special design by an expert.

On the other hand, the areas and the number of applications, are increasing rapidly. Application fields like medicine, biology or finance are boom areas"

and pose a number of new and interesting questions.

The development of system identification now is more driven by demand from applications than by developments in theory, i.e. there is "demand pull" rather than "theory push".

One can also observe an increasing fragmentation corresponding to different data structures, model classes and prior knowledge in different fields of application. The development of theory and methods is also done by different, not very much interacting communities, like econometrics, system- and control theory or signal processing.

System identification is in a certain sense an enabling technology and in many cases not visible for non-experts.

There are still major open problems in system identification, such as

- large parts of identification of nonlinear systems
- Identification of spatio-temporal systems
- Identification for large data sets data and for high dimensional time series
- Improved model selection and regularization procedures
- Further automatization
- Hybrid procedures
- The use of symbolic computaion

Summing up, system identification is still an interesting area, in particular new applications pose new challenges. The field has shifting boundaries and the question arises, whether in the future there will be still a substantial common body of theory and methods. Besides the danger of fragmentation, for certain parts of the field, there is also the danger of becoming selfreferential and not relevant for applications.

References

[1] H. Akaike, Canonical Correlations Analysis of Time Series and the Use of an Information Criterion, in: R.H. Mehra and D.G. Lainiotis, eds., *System Identification: Advances and Case Studies,* Academic Press, New York, 27 – 96, 1976.

[2] D. Bauer, Comparing the CCA Subspace Method to Pseudo Maximum Likelihood Methods in the Case of No Exogenous Inputs, *Journal of Time Series Analysis*, 26, 631-668, 2005.

[3] A. Chiuso and G. Picci, The Asymptotic Variance of Subspace Estimates, *Journal of Econometrics,* 118*,* 292-312*,* 2003.

[4] H. Cramer, *Mathematical Methods of Statistics,* Princeton University Press, Princeton, 1946.

[5] M. Deistler, *The Properties of the Parametrization of ARMAX Systems and Their Relevance for Structural Estimation and Dynamic Specification, Econometrica* 51, 1983, 1187 – 1208.

[6] M. Deistler, *System Identification General Aspects and Structure,* in G. Goodwin (ed.) *System Identification, and, Adaptive Control* (Festschrift for B.D.O. Anderson), Springer, London, 2001, 3 – 26.

[7] M. Deistler, *Linear Models for Multivariate Time Series Analysis,*, in: "Handbook of Time Series Analysis", Matthias Wintherhalder, Bjoern Schelter, Jens Timmer, eds., Wiley-VCH, Berlin, 2006, 283 – 306.

[8] M. Deistler, W. Dunsmuir and E.J. Hannan, Vector Linear Time Series Models: Corrections and Extensions, *Adv. Appl. Probab.,* 10,1978, 360 – 372.

[9] M. Deistler, K. Peternell and W. Scherrer, Consistency and Relative Efficiency of Subspace Methods, *Automatica,* 31, 1865-1875, 1995.

[10] M. Deistler and B.M. Poetscher, The Behaviour of the Likelihood Function for ARMA Models, *Adv. Appl. Probab.,* 16, 1984, 843 – 865.

[11] W. Dunsmuir and E.J. Hannan, Vector linear time series models, *Adv. Appl. Probab.,* 8, 1976, 339 – 364.

[12] E.J.Hannan, The Asymptotic Theory of Linear Time Series Models, *J. Appl. Probab.* 10, 1973, 130 – 145.

[13] E.J.Hannan, Estimating the dimension of a linear system, *J. Multivariate Anal.* 11, 459-473, 1981.

[14] E.J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems,* Wiley, New York, 1988.

[15] W.E. Larimore, System Identification, Reduced Order Filters and Modeling via Canonical Variate Analysis, in: H.S. Rao and P. Dorato, eds., *Proceedings of the* 1983 *American Control Conference,* 445-451, 1983.

[16] H. Leeb and B.M. Poetscher, Model Selection and Inference: Facts and Fiction, *Econometric Theory*, 21, 21 – 59.

[17] L. Ljung, *System Identification. Theory for the User,* Prentice Hall; 2nd ed., 1998.

[18] T. McKelvey and A. Helmersson, System Identification using an Over-Parametrized Model Class - Improving the Optimization Algorithm, in: *Proceedings* 36*th IEEE Conference on Decision and Control*, San Diego, USA, 2984-2989, 1997.

[19] B.M. Poetscher and I. Prucha, *Dynamic Nonlinear Econometric Models: Asymptotic Theory*, Springer, New York, 1993

[20] T. Ribarits, M. Deistler and McKelvey, An Analysis of the Parametrization by Data Driven Local Coordinates for Multivariable Linear Systems, *Automatica*, 40, 789-803, 2004.

[21] G. Schwarz, Estimating the dimension of a model, *Ann. Statist.*, 6, 461-464, 1978.

[22] R. Shibata, An optimal autoregressive spectral estimate, *Ann. Statist.*, 9, 300-306, 1981.

[23] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems: Theory, Implementation, Applications,* Kluwer Academic Press, Boston, 1996.

[24] A. Wald, Note on the consistency of the maximum likelihood estimate, *Ann. Math. Stast.,* 20, $595 - 601$, 1949.

[25] J.C. Willems, Thoughts on System Identification, in: B.A. Francis and J.C. Willems, eds. *Control of Uncertain Systems: Modelling, Approximation and Design (Festschrift for K. Glover)*, Springer Lecture Notes in Control and Information Sciences, 2006.